A Relational Database for Agronomic and Genealogical Sugarcane Data: An Adaptable Prototype

Eric T. Stafne, James S. Brown, and James M. Shine, Jr.*

ABSTRACT

A comprehensive relational database has been created at the USDA-ARS Canal Point Sugarcane Field Station to facilitate the entry and retrieval of data for the breeding program using Microsoft Access 2000. This software is readily available and easily adaptable to a wide variety of breeding programs. A relational database provides an efficient and powerful way to store, retrieve, manipulate, query, and report data in a multiuser environment. Data entry is performed through a series of self-explanatory forms. Once entered, data can be accessed and queried through a local area network (LAN). Data from the seedling stage (true seed planting), Stage I and Stage II (the first two clonally propagated selection stages at Canal Point), and Stage IV (the final selection stage before cultivar release) of the breeding program have currently been included in the database. The database also includes the Canal Point clonal collection inventory, crossing information, seed (fuzz) inventory, and pedigree tracking. Future plans include incorporation of data from Stage III (the next to last clonally propagated selection stage), pathology data, and access through the Canal Point Internet web site (www.canalpoint.sugarcane.usda.gov).

The release of agronomically elite cultivars is an important goal of the USDA-ARS Sugarcane Field Station at Canal Point, FL. The selection cycle from crossing to variety release is approximately 10 yr. Sugarcane (*Saccharum officinarum* L.) breeding and clonal development occurs in several stages. The initial phase involves crossing of parental clones to produce true seed followed by germination testing, storage, and planting to produce approximately 50 000 to 100 000 seedlings annually for evaluation. Seedlings are selected at an intensity of 10 to 15% and subsequently progress to Stage I (of a four-stage testing program). Individual selections are planted by stool (or layering) and then visually selected for agronomic type (i.e., stalk height, diameter, and erectness) and disease resistance several months later. Selected clones are coded with a *CP* number designating their origin as Canal Point, FL. Stage II follows approximately 1 yr later. Clones are selected from larger, multistool plots. Aside from the previously mentioned selection criteria, cane yield and percent sucrose are measured in this stage. Approximately 140 clones are advanced to Stage III, in which clones are grown in doubly replicated plots at four locations, evaluated for disease susceptibility and agronomic character, and measured again for cane and sucrose yield. Ten to 12 clones are advanced for Stage IV testing after both plant cane and first-ratoon seasons. Stage IV is the final testing phase before release and is evaluated over three crop seasons: plant cane, first ratoon, and second ratoon. Extensive pathology data from inoculated tests and natural infection are accumulated during Stage IV evaluation. The final selections are increased vegetatively and usually grown at 10 different cooperative grower farms around the Everglades Agricultural Area. This program has averaged the release of about one clone per year. A thorough description of sugarcane breeding is available in Dunckelman and Legendre (1982). Tai and Miller (1989) gave a detailed account of the Canal Point selection program.

Computerization of sugarcane records and creation of programs for routine breeding tasks and data calculation began around 1960 (Meyer et al., 1974). Miller and James (1973) reported the use of a programmable calculator in the Florida breeding program. Wu (1987) gave a detailed overview of software for management of the sugarcane breeding program in Hawaii and a review of computer use in sugarcane breeding at the time. Morris et al. (1982) described the development and implementation of a computerized data management system for the Louisiana sugarcane breeding programs at Louisiana State University and for the USDA-ARS program at Houma, LA. This system was originally designed to predict progeny performance, manage seed inventory and progeny testing data, and determine experimental variety performance. One of the difficulties encountered by these researchers was educating programmers about field procedures, specific phases of selection, and related sugarcane terminology. Current computer-based programs enable plant breeders to build their own databases with less investment in hardware and programming, thus optimizing time and effort in breeding endeavors.

Breeding programs, like all large-scale research programs, inherently produce large amounts of data that must be summarized in various ways. A relational database system combined with modern statistical analysis and graphical presentation tools is the most effective system for storing, retrieving, summarizing, and presenting acquired data. All of the data collected at each stage throughout the 10 or more years of evaluation must be available to decision makers involved in cultivar release to obtain maximum benefit. In order to make informed selection decisions, researchers and industry leaders need to have access to the data used in the breeding program. Data are initially manipulated and/or summarized through queries to provide the desired output without excess information that can hamper interpretation. More thorough statistical analyses and graphical presentations may be required following initial summar-

E.T. Stafne, Dep. of Hortic., Univ. of Arkansas, 316 Plant Sci., Fayetteville, AR 72701; J.S. Brown, USDA-ARS, Subtropical Hortic. Res. Stn., 13601 Old Cutler Rd., Miami, FL 33158; and J.M. Shine, Jr., Sugarcane Growers Coop. of Florida, P.O. Box 666, Belle Glade, FL 33430. Received 11 Sept. 2000. *Corresponding author (jmshine@scgc.org).

Published in Agron. J. 93:923–928 (2001).

Abbreviations: MS, Microsoft.

Table 1. Examples of table structure and associated parameters used in the Canal Point breeding program database.

Table name	Field name	Type	Indexed	Field description
Tblcrosslist (crossing information, updated once per year)				
	ID	AutoNumber	Yes	Computer-generated identification
	Year	Number	Yes	Year of cross
	Crossno	Text	Yes	Number of cross
	Crossdate	Text	No	Date cross was made
	Source	Text	Yes	Source of crossing material
	Female	Text	Yes	Female parent
	Male	Text	Yes	Male parent
	Dist	Text	Yes	Distribution of final cross
	Ttas	Number	No	Total number of tassels produced
	Gmfuzz	Number	No	Grams of fuzz produced
	Seedp	Number	No	Seed per gram produced
	Binno	Text	No	Storage location
	Project	Text	Yes	Project identification
Tblfuzz (fuzz and seed inventory information)				
	ID	AutoNumber	Yes	Computer-generated identification
	Crossno	Text	Yes	Number of cross
	Invdate	Text	No	Date fuzz put in inventory
	Binno	Text	Yes	Storage location
	Gmfuzz	Number	No	Grams of fuzz produced
	Seedpgrm	Number	No	Seed per gram produced
	Germdate	Text	No	Date of germination (in test)
	Female	Text	Yes	Female parent
	Male	Text	Yes	Male parent
	Project	Text	Yes	Project identification
Tblcount2 (seedling planting and selection information)				
	ID	AutoNumber	Yes	Computer-generated identification
	Record	Number	No	Field identification number
	Year	Number	No	Year of planting
	Side	Text	No	Field orientation
	Block	Number	No	Block of planting
	Row	Number	No	Row of planting
	Direction	Text	No	Row orientation
	Crossno	Text	Yes	Number of cross
	L Side	Number	No	Number of plants, left row
	R Side	Number	No	Number of plants, right row
	Selected	Number	No	Number of plants selected
	Comments	Text	No	Observations in field
Tblworld1 (world collection information)				
	ID	AutoNumber	Yes	Computer-generated identification
	CPno	Text	Yes	Number given to Canal Point clone
	Impno	Number	Yes	Import number
	Location	Text	Yes	Where clone is held
	Block	Text	Yes	Block in field where planted
	Tier	Number	Yes	Tier in field where planted
	Rowno	Number	Yes	Row in field where planted
	PIno	Text	Yes	Plant introduction number

ies. The database provides safe data storage that can be secured and shared over a local area network (LAN) and linked to the Internet for access by growers and other industry members. Sugarcane breeding has several unique properties, and no commercial software package is currently available. Therefore, we chose to develop a program that incorporates all relevant aspects of the breeding program. A relational database such as this one is designed to centralize storage of a main copy of experimental results though many backup copies and data subsets exist as well.

SOFTWARE SPECIFICATIONS

Microsoft (MS) Access 2000[1] (Microsoft Corp., Redmond, WA) software was used to construct the database and provide a user-friendly interface for data entry. The database requires a 486 personal computer or better using at least a 75 MHz processor with a Windows-based operating system (Windows 95, Windows NT Workstation 4.0 or later). The system requires at least 24 megabytes (Windows 95) or 40 megabytes (Windows NT) of random access memory (RAM), a monitor of video graphics array (VGA) resolution or better, and a mouse.

[1] Mention of trade names or commercial products in this paper is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA.

DATABASE DESIGN

In the early 1990's, a disk operating system (DOS) based database using Knowledgeman/2 (mdbs, Lafayette, IN) software was implemented at the Canal Point station by James Shine, Jr., Florida Sugar Cane League. This program was used to manage data from the crossing program, Stage I, Stage II, seed inventory and shipment, and germplasm collection. The program code was not Y2K compliant or network accessible, necessitating revision of the program. Several new features were also desired; therefore, reconstruction of the database was done in MS Access 2000.

Several tables (similar to spreadsheets) from the previous database were converted to MS Excel using DBMS/

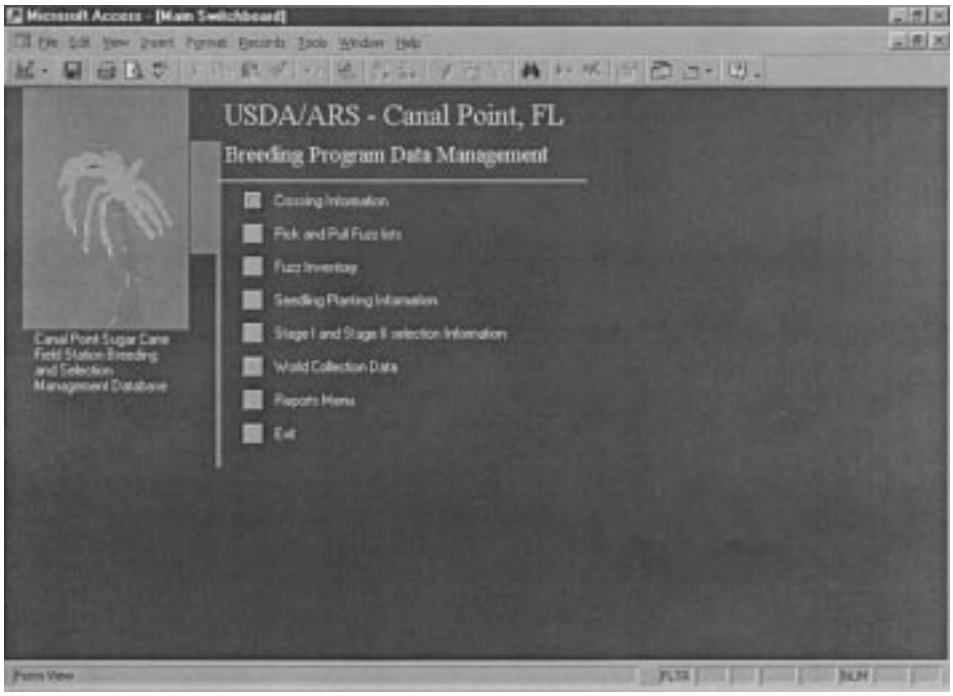


Fig. 1. The main switchboard for the breeding database of Canal Point field station, which facilitates navigation throughout the database.

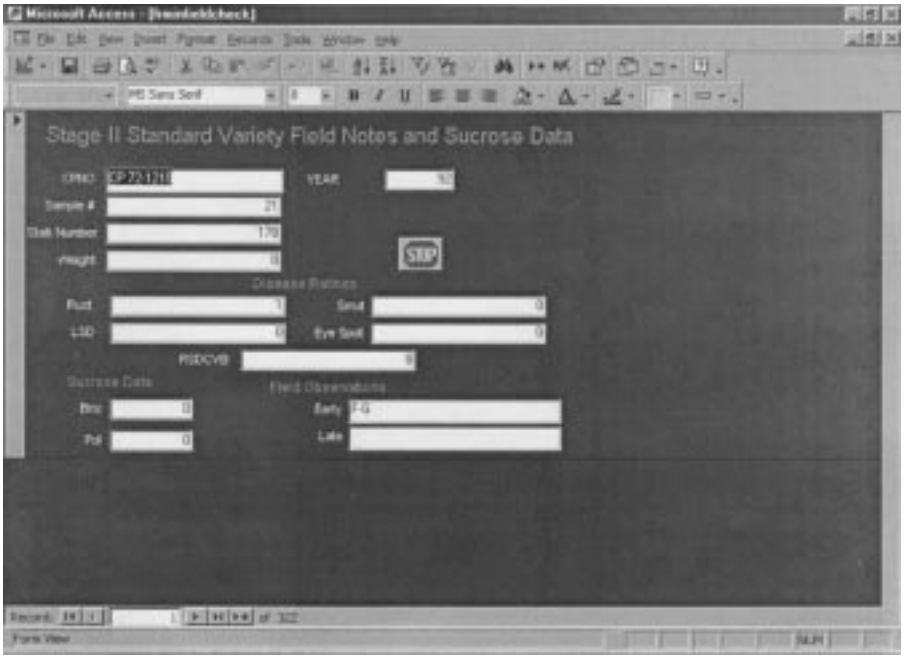
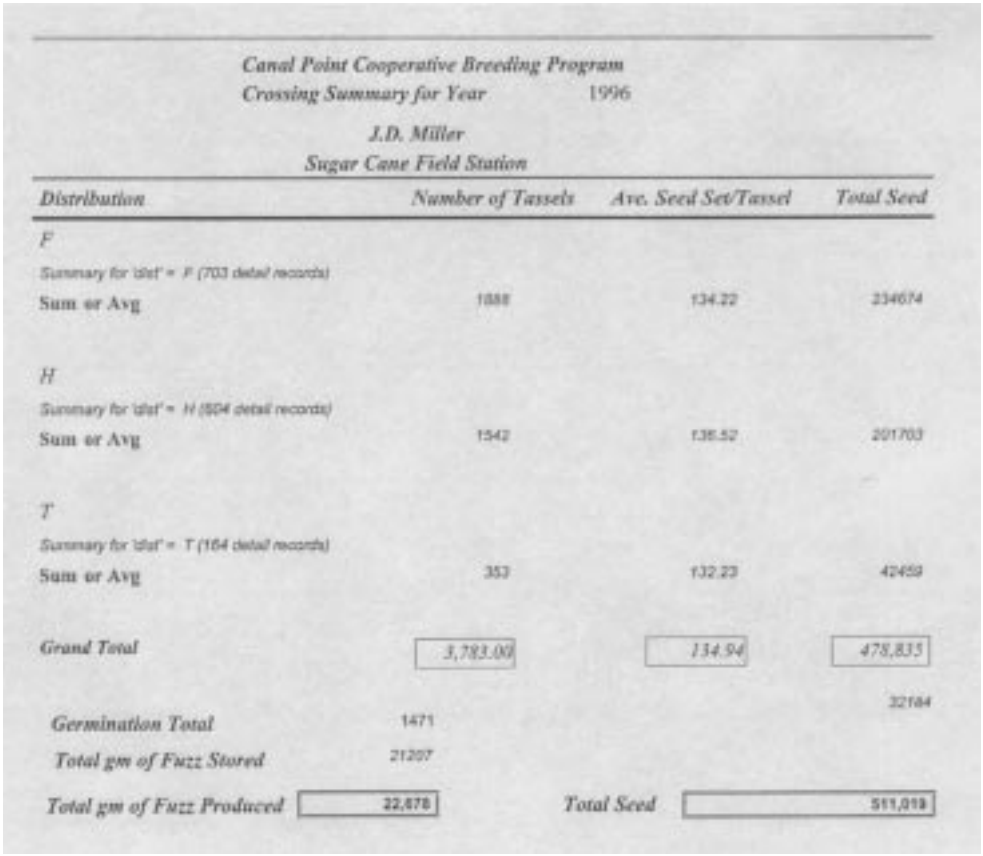


Fig. 2. An example of a form where data is entered in the breeding database of the Canal Point field station.

Copy 6.0 (Conceptual Software, Houston, TX) and subsequently imported into MS Access 2000. All queries, forms, and reports were constructed following the code written in the previous database to maintain familiarity for users of the prior program. The code was mostly generated using *wizards*, a very useful option in the MS Access 2000 software package that eliminates most of the need for tedious code writing. *Design view* is another method of constructing tables, queries, forms, and reports. This approach allows more flexibility than the wizard option but also requires more in-depth knowledge of programmatic components. Introductory knowledge of Structured Query Language (SQL) and Visual Basic commands (VBA) were used for specialized assignments. Additional forms and reports were designed to enhance the options of the new database.

DATABASE STRUCTURE

A database designed in MS Access consists of a few main components: tables (similar to spreadsheets or field ledgers), queries, switchboards, forms, and reports.



Canal Point Cooperative Breeding Program
Crossing Summary for Year 1996
J.D. Miller
Sugar Cane Field Station

Distribution	*Number of Tassels*	*Ave. Seed Set/Tassel*	*Total Seed*
F			
Summary for 'dist' = F (703 detail records)			
Sum or Avg	1688	134.22	234674
H			
Summary for 'dist' = H (804 detail records)			
Sum or Avg	1542	138.52	201703
T			
Summary for 'dist' = T (164 detail records)			
Sum or Avg	353	132.23	42459
Grand Total	3,783.00	134.94	478,835
			32184
Germination Total	1471		
Total gm of Fuzz Stored	21207		
Total gm of Fuzz Produced	22,678	*Total Seed*	511,019

Fig. 3. A summary report output through the breeding database of the Canal Point field station. This report was produced with the *wizard* option.

Another important step needed to accomplish full functionality is to create *relationships*. Relationships are links between indexed fields (columns within a table) in a table or query. Different relationship types exist, such as one-to-many or many-to-one. Indexing is a means of cataloguing the data in a particular field. This expedites the sorting process in a query.

Tables are the starting point for any database. A table may contain all of the data for a particular program, or it may be split into several tables, which can then be linked through a common indexed column. Each table has a unique structure, with field (column) headings and the data that fits under those fields, as in a spreadsheet. Examples of table structure with field names, data type, and indexing information can be seen in Table 1.

Queries use single or multiple tables. If multiple tables are used within a query, a relationship must be established between common indexed fields in each table. Queries are the essential sorting method of the mass amounts of data that are stored within the database. One can invoke a prompt query, which asks the user for a defining factor (e.g., year, cross number, and female parent), or a report for only one certain trait (such as a disease or a yield limit) can be constructed. Formulae, sorting, and selection criteria can be incorporated to optimize the search for data of interest.

A switchboard menu allows the user to choose access to forms for data input or reports for data output. Because the switchboard is the primary interface for users, only forms and reports are accessed through the switchboard. Nonetheless, users are provided with a great deal of navigational choice through this approach (Fig. 1).

Forms can be created in two distinct ways: (i) by using a wizard that takes the database creator through a step-by-step process and (ii) by using the design view interface, which allows for greater autonomy of structure but is not as clearly self evident. Once forms are created, they are the subsequent modes of entry for their respective data (Fig. 2). Data are passed from the forms to be stored in a designated table. It is important that forms be created with ease of use in mind because difficult-to-use forms tend to be avoided by those without well-developed computer skills.

A report is where data are output in a fashion that is readily interpretable. It also provides a hard copy that can be saved for archiving. Summary reports can be created to condense data into more easily interpreted output (Fig. 3). At the other extreme, a report can contain every detail of data in a table, or if linked, several tables.

SPECIFIC DATA OF IMPORTANCE

Several types of data are entered throughout the four stages of a breeding program for sugarcane cultivars. Some of the data types are consistently used from one stage to the next. Sucrose and yield data are obtained and/or calculated from Stage II through Stage IV. These data include Brix and polarimeter readings, stalk weights, and stalk population counts. The data are then used to

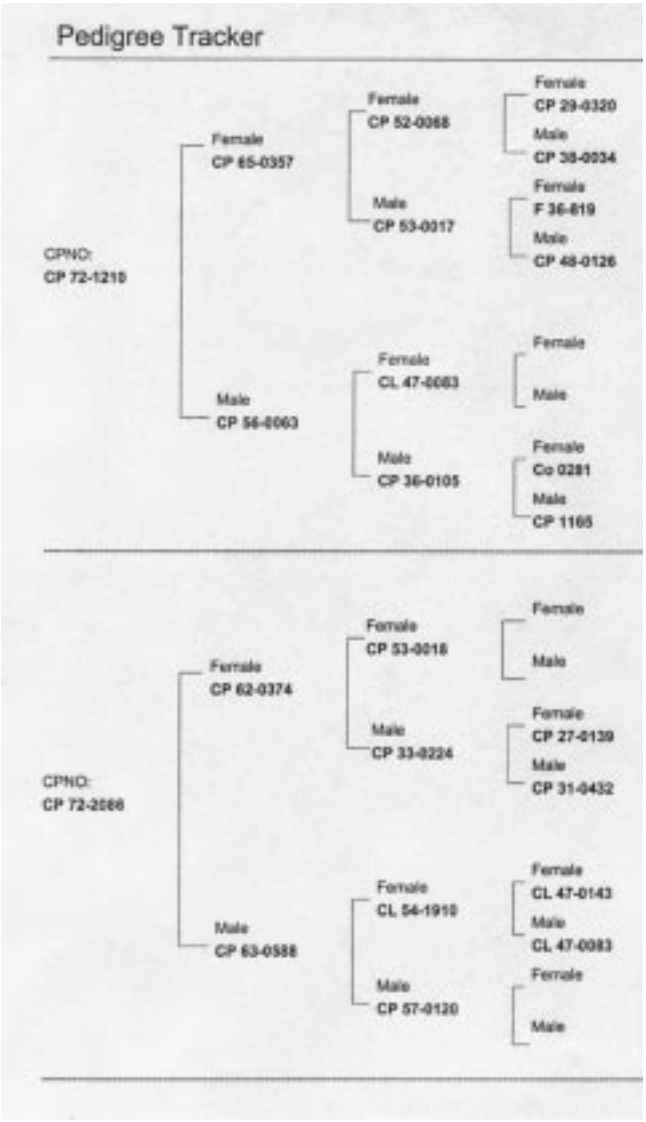


Fig. 4. Pedigree examples of two of the most important clones produced from the breeding database of the Canal Point field station.

calculate sucrose, total recoverable sugar, tonnes of cane per acre, sugar per hectare, and sugar per tonne. Having similar data fields across stages allows the user to view the data of a particular sucrose or yield component for a single clone through all selection stages.

Other stages in the selection process have data that are unique to that specific stage. Data taken during the crossing stage consist of number of tassels, average seed set per tassel, and total seed produced per cross. Seed (fuzz) is collected and tested for germination rate. The number of grams and viable seed per gram are entered in the database and used for inventory control and planting plans. Thirteen years of crossing data are currently stored in the database.

Another unique data set, the collection of sugarcane clones from around the world, is maintained at Canal Point. The Canal Point station—in conjunction with the USDA-ARS Subtropical Horticultural Research Station, Miami, FL—is responsible for much of the world collection of sugarcane and related grasses. Therefore, the station keeps many types of foreign cane in the field as a collection for global use. Data recorded in the database for these sugarcane clones include collection source, location in field, plant introduction and import number, and taxonomic classification. By keeping this information current, collaboration can occur between the Canal Point station and others worldwide.

SPECIAL FEATURE

A pedigree tracker (Fig. 4) was developed to view ancestral information for all sugarcane cultivars currently in the database. This is performed through a series of complex queries when one chooses a clone or clones for which lineage is desired. The output is presented as a report in the form of a dendrogram or family tree of the pedigree. Currently lineages are traced back to the great-grandparental generation, provided the ancestral history to that point is available. Blanks are placed in appropriate pedigree gaps where parentage is unknown. This feature facilitates crossing and release decisions as well as possible tracing of heritable traits from one generation to the next. Wu (1987) reported a similar pedigree program used by the Hawaiian Sugar Planters' Association.

FUTURE ADDITIONS AND APPLICATION

This database is an evolving project with several additions under development. Stage III and pathology data will be added to the database in the near future. These two types of data are more difficult to incorporate into the current database system because of previous data-recording methods. These data modules will allow more extensive and complete queries and increase precision in the variety selection process because historical data will be readily available for any clone.

A cross–progeny performance predictor, similar to that described in Morris et al. (1982), is another desirable component that will be incorporated into the Canal Point database. This will improve efficiency of genetic improvement in the crossing program. It will not be used to narrow the genetic base of sugarcane crosses, but to provide a better predictor of crosses that will produce commercially viable and/or disease-resistant progeny.

Another possibility is developing standardized reporting procedures through computerized tagging and field books. The tedious task of writing and rewriting tags can be replaced by using computer-generated tags. Though initially more expensive, use of computer-generated tags will save time and effort. Also, if weather-resistant material is utilized, then the tags can be reused in subsequent years. This approach has been tested in sugarcane seedling plantings and found to be quite successful. Field books can be printed out in the same format year after year, thus fostering familiarity among all personnel. These features will be added in future revisions of the program.

Future plans also include data access through the Canal Point web site (www.canalpoint.sugarcane.usda.gov). Outside access will be limited mostly to reports.

However, data entry over the Internet with annotation may eventually become appropriate. This could play a significant role in the functionality of the database for local and remote users worldwide.

Plant breeders must be able to access and query data from many different perspectives. The secure environment of the database ensures immediate access to results. Therefore, a functional database providing ease of access and query functions is essential for any modern breeding program.

Requests for information concerning this database, including code, can be sent to jmshine@scgc.org or jdmiller@saa.ars.usda.gov.

ACKNOWLEDGMENTS

We wish to acknowledge the helpful reviews and suggestions of Dr. Steve Larson and Dr. Kenneth Gravois.

REFERENCES

Dunckelman, P.H., and B.L. Legendre. 1982. Guide to sugarcane breeding in the temperate zone. Agric. Rev. and Manuals 22. USDA-ARS, New Orleans, LA.

Meyer, H.K., D.J. Heinz, N. Lawrence, E. Kimura, and S.L. Ladd. 1974. Computer processing of sugarcane yield, breeding, and selection records. Proc. Int. Soc. Sugar-Cane Technol. 15:24–35.

Miller, J.D., and N.I. James. 1973. The use of programming calculators in sugarcane breeding. Proc. Am. Soc. Sugar Cane Technologists 2:49–52.

Morris, D.D., R.D., Breaux, and F.A. Martin. 1982. The use of computers in managing sugarcane breeding data in Louisiana. Proc. Inter-Am. Sugarcase Seminar 3:58–65.

Tai, P.Y.P., and J.D. Miller. 1989. Family performance at early stages of selection and frequency of superior clones from cross among Canal Point cultivars of sugarcane. J. Am. Soc. Sugar Cane Technologists 9:62–70.

Wu, K.K. 1987. Computer applications in sugarcane improvement. p. 543–558. *In* D.J. Heinz (ed.) Developments in crop science 11: Sugarcane improvement through breeding. Elsevier, Amsterdam.